Developing a Training Data Set for the CCL3 Classification Process

Zeno Bain[1], Thomas Carpenter[1], Joyce Donohue[1], Michael Messner[1], Wynne Miller[1], Yvette Selby-Mohamadu[1], Frank Letkiewicz[2], JoAnne Shatkin[2], and George Hallberg[2]

[1] USEPA, Office of Water, Washington, DC, USA.

[2] The Cadmus Group, Watertown, MA, USA.

BACKGROUND

EPA plans to use classification rules and models to help identify contaminants for its Contaminant Candidate List (CCL). Classification Models (described in a subsequent poster DS-3) are algorithms that derive mathematical relationships among input variables and class-association (output). For Contaminant Candidate List 3 Process, these models were used to develop relationship between attribute scores and list/no list decisions.

A training data set (TDS) for the purposes of CCL 3 is the set of data used to train or teach classification models to mimic expert list-no list decisions. The TDS is a large number of chemical contaminants that has two components for each:

- Contaminant Attribute Scores (Using formal Attribute Scoring Protocols (ASPs))
- List/No List Decisions (Formed through Expert Consensus)

CONTAMINANT ATTRIBUTE SCORES

The four attributes (described in a previous poster DS-1) are organized by health effects and occurrence:

Health Effects

Potency – amount of a contaminant required to cause an adverse health effect

Severity – the adverse health effect observed at the lowest toxic dose for a contaminant

Occurrence

Magnitude – concentration at which a contaminant is known or anticipated to occur in drinking water

Prevalence – how commonly a contaminant occurs or has the potential to occur in drinking water

LIST/NO LIST DECISIONS

List/no list decisions are assigned to one of four categories:

- Not List (1)
- Not List? (2)
- List? (3)
- List (4)

Note: Also available are average classifications, such as 3.5 when half the experts assigned the contaminant to List and half assigned it to List?

PRINCIPLES and OBJECTIVES

A TDS should:

- have contaminants that represent a range of outcomes and decisions likely to be encountered in developing a CCL
- include a variety of input data ensuring adequate coverage of attribute scores and combinations in order to train the classification models
- contain enough contaminants to adequately train the classification models being considered
- have input information and decisions that adequately reflect the list-no list decision-making process to ensure consistency and confidence in outcomes
- incorporate decisions made by experts that identify priority contaminants for consideration and protects public health

EXAMPLE TDS CONTAMINANTS

The following examples are possible scores and decision results of TDS contaminants that are inputs to the classification models:

Contaminant	INPUT ATTRIBUTES				DECISION
Number	Potency	Severity	Prevalence	Magnitude	L/NL
30	10	9	9	10	L = 4
36	6	9	1	1	L? = 3
40	4	6	3	3	NL? = 2
43	4	3	3	2	NL = 1

DEVELOPING THE TDS

Attribute scoring protocols (ASPs) were developed to consistently score each of the four attributes based on the range and hierarchy of available data. Each contaminant of the TDS was scored using these ASPs. As the example TDS contaminants above demonstrate, experts made decisions using only the four attribute scores, without contaminant names identified. Separately, experts made decisions using the range of available data, without attributes scores and the contaminant names identified. Both decisions were compared for consistency and used to refine the ASPs.

DEVELOPING THE TDS continued..

Once the final ASPs were developed, an expert panel made individual decisions on all of the TDS contaminants using only the four attribute scores, without the contaminant names identified. For both decisions based on the data and attribute scores, experts met in a group to discuss decisions, learn from each other, and develop consensus. Examples of consensus decisions and individual expert decisions are displayed below for two example contaminants:

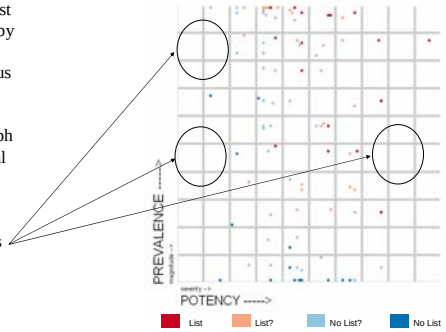
Contaminant #	INPUT ATTRIBUTES				Blinded Attribute Scores Decisions						INPUT DECISION	
	Pot.	Sev.	Prev.	Mag.	Expert1	Expert2	Expert3	Expert4	Expert5	Expert6	List=4 Mean	Mean L/NL
153	6	9	1	4	NL?	NL	NL	L?	L?	L?	2.17	NL?
154	8	8	8	2	L	L	L?	L	L?	L	3.67	L

FIRST 101 CONTAMINANTS OF THE TDS

This graph displays the first 101 contaminants plotted by the scores of the four attributes and the consensus decisions.

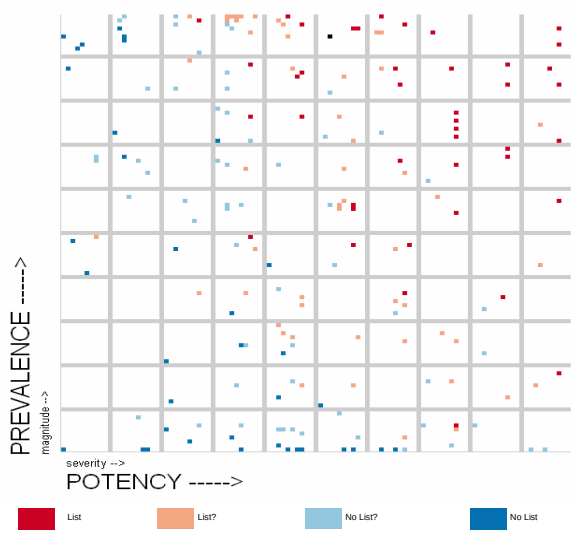
The boundaries of the graph show the range of potential attribute scores.

Notice some of the visible gaps in the attribute scores coverage for the first 101 contaminants.



FINAL 202 CONTAMINANTS OF THE TDS

The performance of the classification models using the initial TDS gave an indication that the TDS was not adequately covering the range of attribute scores. To cover this space contaminants were added to the TDS. The additional set of 101 "artificial" contaminants were developed with specific combinations of attribute scores to fill in gaps in the attribute scores coverage and improve the performance of the model. Model results are displayed on a subsequent poster (DS-3).



REFERENCES

Classifying Drinking Water Contaminants for Regulatory Consideration, National Research Council, Committee on Drinking Water Contaminants, NRC Press, 2001.

National Drinking Water Advisory Council Report on the CCL Classification Process, 2004.

Graph colors based on www.ColorBrewer.org, by Cynthia A. Brewer, Penn State Univ.